

Ch 2 Data Analytics

What is a hypothesis?

Example:

Listening to music for more than an hour will lead to lower test scores than studying without music because you are not completely focused on the topic.

Definition (from VCAA)

- ▶ A hypothesis typically describes a cause and effect.
 - ▶ It should include at least one prediction (that does not arise from another explanation).
 - ▶ It is a statement (or prediction) about the relationship between two or more variables. That is, a change in one variable will result in a change in the other.
- It must be testable.

Breaking down the definition.

- ▶ A hypothesis typically describes a **cause** and **effect**.



<https://www.youtube.com/watch?v=JXtDC7hNB3gc>

Cause

Listening to music for more than an hour will lead to **lower test scores** **Effect** than studying without music because you are not completely focused on the topic.

Breaking down the definition.

- ▶ It should include at least one prediction (that does not arise from another explanation).
- ▶ **Prediction** is that test scores will change if listening to music is decreased. The **reasoning** is that the student will be more focused.

Breaking down the definition.

- ▶ It is a statement (or prediction) about the relationship between two or more variables. That is a change in one variable will result in a change in the other.
- ▶ **Independent variable:** the variable in the experiment that is changed or manipulated eg: listening to music.
- ▶ **Dependent variable:** the variable in an experiment that responds to change eg: lower test scores.

Breaking down the definition.

- ▶ It must be testable.
- ▶ Test scores can be measured.

Try this example

When boxes of chocolates are given to teachers, the amount of homework decreases because the teacher will want to please the students.

Identify the following:

- › Independent variable (what are we changing).
- › Dependent variable (what will change as a result).
- › Prediction (what will happen).
- › Reasoning (why will it happen).
- › Is there a cause & effect?
- › Is the hypothesis testable?

Data sources

- ▶ Treat data from unidentified sources with care and take steps to authenticate the data (second-sourcing).
- ▶ Establish a "chain of custody" – from creator to you.
- ▶ A reputable data source that has authority will more likely provide high-quality data.

Data sources

Because there is so much data available (especially online) it is vital the data selected is appropriate.

A combination of primary & secondary sources can assist in having the data available to solve an information problem (such as supporting or refuting a hypothesis).

Data integrity is very important. This will be discussed later.

Data sources

There are hundreds of data sources available online in Victoria and Australia alone.

- ▶ Libraries, whether local, state or national.
- ▶ Government departments and agencies (BOM, ABS etc.).
- ▶ Private organisations such as businesses or societies.

Acquiring data

Data can be acquired using a variety of methods including:

- ▶ observation
- ▶ interviews
- ▶ Surveys & questionnaires
- ▶ Querying existing databases

However to acquire data you first need to know how to ask questions.

Acquiring data Asking questions

Closed questions

- ▶ Usually found in surveys & questionnaires.
- ▶ Limited choice of responses, so can use tick boxes, radio buttons etc.
- ▶ Quicker & easier to collect and collate data.

Acquiring data Asking questions

Open questions

- ▶ Usually found in interviews.
- ▶ No limits on responses. More able to give opinions and ask follow-up questions.
- ▶ Responses are more detailed and unique.
- ▶ More time consuming to collect and difficult to collate data.

Acquiring data Asking questions

Faults in questions

Some questions can be flawed which impacts on the quality of the data collected.

- ▶ Loaded questions can disguise a hidden agenda. "Why do you think immigrants are involved in so much crime in Australia".
- ▶ Leading questions can be biased in certain directions. "Do you think murderers should get life imprisonment or the death penalty"?
- ▶ Closed questions can often be biased in this way through the options they allow.

Acquiring data Interviews

- ▶ Real time, personal interactions.
- ▶ Allows the questioning to be shaped by the responses given.
- ▶ Can pick up on subjects body language.
- ▶ Can be conducted using video technology if required.
- ▶ Record if possible.
- ▶ Important that interviewer is well prepared and that subject is comfortable and relaxed.

Acquiring data Observations

- ▶ Observing behaviour can be informative because people are not influenced by the data collector (covert v overt).
- ▶ Can pick up on non-verbal behaviour (eg: body language).
- ▶ May have electronic aspects of measuring (eg: audit trails and log-ins).
- ▶ Costly in terms of time & money.
- ▶ Susceptible to observer bias.

Acquiring data Surveys & questionnaires

- ▶ Often used as starting point for research when little is known about the subject.
- ▶ Used with larger groups often in a written (or online) format.
- ▶ A great deal of data can be gathered relatively quickly and cheaply.
- ▶ Don't allow for clarification or follow-up with the subject.
- ▶ Easy for respondents to be untruthful, leave out questions or respond in the way they think the data collector would expect.

Acquiring data Querying resources

- ▶ Often the data you require will be contained in large, complex data sets, so an ability to query (filter) the data will be vital.
- ▶ Usually built-in services are available to do this.
- ▶ Query by example (QBE) - type in a description of the data you are seeking (eg: Google). See p 105.
- ▶ Structured Query Language (SQL) - common method of requesting data from a database. It actually creates code to search for your query. Your queries in Access were created this way. See p 106.

Acquiring data Referencing data sources

It is important that any data sources you use are acknowledged in your work. We do this to avoid:

- ▶ Claiming other people's work as our own (plagiarism).
- ▶ Readers can find the original source if required.
- ▶ Copyright laws are observed.
- ▶ Moral rights are observed.

Solution specifications

When creating a solution to a problem we need to consider a number of factors which make of the specifications of the solution.

The specifications of a solution consist of:

- ▶ Requirements
- ▶ Constraints
- ▶ Scope

Requirements

What your information needs to do, the qualities it should have and what data is required to create the information.

For example think about the output required (info) by the users. What should it tell them? How are they going to use the info?

When this has been decided then the required data can be identified.

Requirements

Functional requirements: the tasks that the solution should be able to perform, eg: what reports, charts etc are required by the user.

SAT Outcome, that it lets your reach a valid & substantiated conclusion as to whether your hypothesis is supported or refuted

Non-functional requirements: the attributes or qualities of the info, eg: accurate, easy to use, secure etc.

Data requirements: the data selected must be able to produce the required output.

Constraints

The limits and restrictions under which your solution must be produced.

This will usually reduce your freedom of design choice.

Constraints

Constraints will usually fall into five categories:

- ▶ **Economic:** relating to cost & time. How long can you afford to spend on creating the solution?
- ▶ **Technical:** availability of equipment (hardware & software), capacity limitations and security requirements.
- ▶ **Social:** expertise of users.
- ▶ **Legal:** legislation such as copyright & privacy.
- ▶ **Usability:** can intended users operate the solution. Does it serve its intended purpose?
- ▶ Example, could not get permission, de-identify the respondents, ran out of time, could not get a picture you wanted

Scope

States the boundaries or parameters, What the info produced must achieve and what it is not required to achieve.

For example what info does your solution need to produce to support or refute your hypothesis; what will or will not be covered

This should be defined precisely so developers can effectively allocate time & resources to the project.

Scope largely defined by its functional & non-functional requirements

Ch 2 Data Analytics;

What is data?

Data is made up of raw, unprocessed facts & figures.

At the data stage it is not in a useful form capable of solving the solution.

When it has been converted (processed) into a useful form then it can be used (for example to support or refute a hypothesis).

Categories of data

Can be categorised in a number of ways, however the most common classifications are:

- ▶ Primary & secondary
- ▶ Quantitative & qualitative

Primary data

Has not been filtered by interpretation or evaluation (unprocessed state).

Usually collected directly by the user however old data that has never been interpreted can be primary data.

Strengths: specific to needs, can be trusted because know source.

Weaknesses: cost & time.

Secondary data

Collected and **interpreted** by someone else.

Examples include professional researchers & Govt. organisations.

Strengths: cheap & quick to collect, huge datasets available, can be useful to support your own findings.

Weaknesses: not specific, sources may not be reliable, may not be able to access original unprocessed data.

Quantitative data

Concerned with numbers and measurement.

It uses an objective approach (no opinions) & closed questions (limits the response).

Strengths: simple to collect, store and analyse. Can easily be compared to previous results.

Weaknesses: limits the responses and does not give people the chance to give their opinion.

Qualitative data

Expressed in words.

It uses a subjective approach (opinions, personal views & experiences), open questions (allows a range of responses).

Strengths: can be used to gain a wider range of responses. Allows people to elaborate.

Weaknesses: difficult to collect, store and analyse (needs to be encoded to be analysed statistically).

Coding qualitative data

Because it can be difficult to work with it is often simpler to encode the data.

This requires responses to be categorised and then allocated a value (see p 86).

Coding reduces original wordiness using freely chosen summary terms, descriptive coding

Rubrics, a detailed list of descriptive grading criteria that correspond with a code, mark)

Some software tools are available to do this

Acquiring data
Referencing data sources

- ▶ This includes direct quotes as well as ideas, summaries or paraphrasing.
- ▶ Various methods of acknowledgement can be used which we will investigate later (see p 109).

Data integrity

- ▶ It is vital if we are to produce useful information that the data we use can be trusted (ie: it has integrity).
- ▶ GIGO
- ▶ Whether storing, transmitting or archiving data integrity must be maintained.

Data integrity

Factors that influence data integrity include:

- ▶ Timeliness
- ▶ Authenticity
- ▶ Relevance
- ▶ Accuracy

Data integrity

Timeliness

- ▶ Should be processed while current and retrieved without significant delays.
- ▶ Decisions should not be based on outdated data.
- ▶ Relies on capabilities of hardware & software used. Chances of potential delays should be reduced.
- ▶ Age of data should always be checked before using.

Data integrity

Authenticity

- ▶ Will be reliant on the reliability of the source and whether this can be verified.
- ▶ Has the data been corrupted or changed?
- ▶ Can the data be second-sourced or check the original documents.

Data integrity

Relevance

- ▶ This measures how closely a resource corresponds with the person's desire for information.
- ▶ Reasons that data can lose relevance include relating to a different time or place, differences in circumstances or getting off topic.

Data integrity

Accuracy

Two main characteristics of accuracy are:

- ▶ Content (functionality). For example correctness & completeness.
- ▶ Form (appearance). For example clarity and consistency (of format).

Data integrity

Accuracy – Correctness

- ▶ Values are correct.
- ▶ Caused by faulty acquisition methods (including equipment failure, misinterpretation of questions, fraud or vandalism, small sample size, data entry errors and time factors)
- ▶ Remove or repair data from other sources. Use validation techniques.

Data integrity

Accuracy – Completeness

- ▶ No missing data (can be difficult to achieve).
- ▶ May have been lost, deleted or unable to be retrieved.
- ▶ Relevant questions may not have been asked or all questions have not been answered.
- ▶ Refer to original data if possible. Ensure fields are not left empty (existence validation).

Data integrity

Accuracy – Clarity

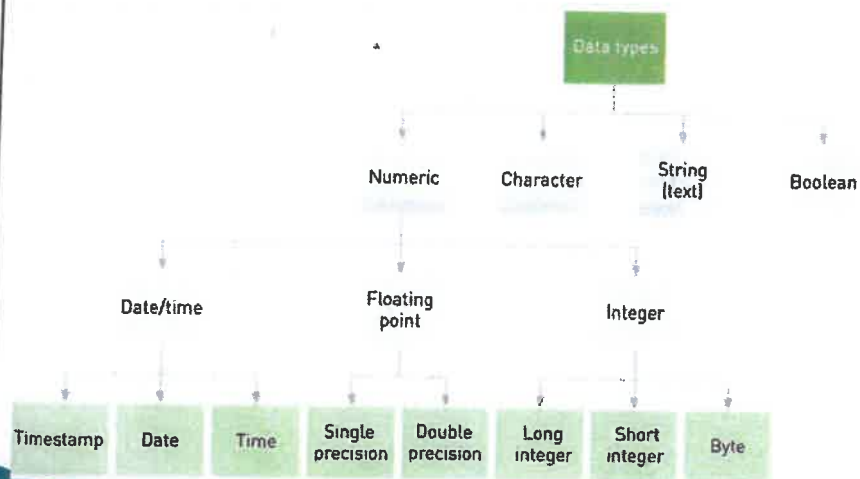
- ▶ Data is clearly presented so it cannot be misinterpreted.
- ▶ Make sure rules for the format of particular data fields (eg: dates).

Data integrity

Accuracy – Consistency

- ▶ The format of the data must be consistent.
- ▶ Inconsistent use of place names (eg: st. or street) or dates (mm/dd/yyyy) or (dd/mm/yy) can make manipulating the data more difficult.
- ▶ Can also be a problem with discrepancies between multiple data sources (eg: different first names – Matthew or Simon).
- ▶ Use validation and data checking between databases. Also ask the same question in different ways to see if the answers are consistent (student/staff/parent surveys).

Data types & data structures



Data types & data structures

- ▶ A Field's data type is separate from its data format
- ▶ A timestamp data type contains both a date and a time of day
- ▶ Integer type cannot store fractional data
- ▶ Spreadsheets guess what data type to use
- ▶ Databases have formal fields, records & tables while spreadsheets do not

Ch 3 Data Analytics, Drawing Conclusions

Legal Requirements

We have discussed earlier in the course the importance of keeping data secure.

Reasons included:

- ▶ Competitive advantage
- ▶ Business reputation & public confidence
- ▶ Ability to function effectively
- ▶ Legal requirements

Legal Requirements

Being in possession of other people's personal data is a heavy responsibility.

Different laws govern how data and information can be collected, stored, communicated, protected used and disclosed.

The key legislation is covered in the presentation.

Privacy legislation

Several federal & state laws govern information privacy.

- ▶ Privacy Act 1988 (federal)
- ▶ Privacy & Data Protection Act 2014 (state)
- ▶ Health Records Act 2001 (state)

Privacy legislation

We will cover this legislation in more detail later in the year. However in simple terms the effect of these acts is to:

- ▶ Restricts how data can be used (only for purpose collected).
- ▶ Organisations responsible to keep the data collected secure.
- ▶ Individuals can request access to data stored about them.

Other legislation

Spam Act 2003 (state)

- ▶ Regulates commercial email.
- ▶ Individuals must consent to receive the email.
- ▶ Commercial messages must identify the sender.
- ▶ Must be able to unsubscribe.
- ▶ Address harvesting software must not be used.
- ▶ Some organisations are exempt (eg: government bodies).

Other legislation

Copyright Act 1968 (federal)

- ▶ Original creative or artistic work is the property of the person who created it (intellectual property) and automatically protected.
- ▶ Protects from unauthorised reproduction, conversion, adaptation, transmission or publication of IP.
- ▶ This is why referencing sources will be important in the SAT.

Other legislation

Charter of Human Rights and Responsibilities Act 2006 (state)

- ▶ Protects an individuals rights of privacy, reputation and freedom of belief and expression
- ▶ Some conflict between the rights of the individual and rights of society are covered in other legislation (eg: racial vilification).

Project Management

A project is an activity which has:

- ▶ a clearly defined purpose
- ▶ a finite lifetime (set start & finish dates)
- ▶ a finite budget
- ▶ a number of interdependent tasks.

Project Management

A number of projects are undertaken in the ICT field, including building or changing information systems.

Therefore it is important to plan for the project to restrict the impact on an organisations operations and finances.

Project Management

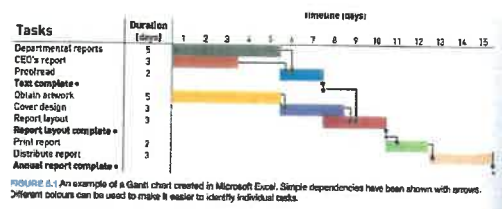
Definition:

Project management involves the planning, organising, and monitoring a project in order for it to be completed on time and within budget.

PM: Gantt Chart

Gantt Chart

- A graphical representation of the project. It indicates the timing and duration of the tasks; can also show milestones and task dependencies



Project Management

Some concepts:

Milestones – represent the achievement of a significant stage in a project. This allows accurate monitoring of progress.

Dependencies – as tasks are interdependent they must be completed in a specific order. A dependent task cannot be commenced until the predecessor has been completed.

Project Management

Processes:

1. Task identification – break down into manageable tasks.
2. Sequencing – work out the dependencies.
3. Time allocation – give each task a time allocation.
4. Documentation – create a Gantt chart (p 138).

Managing data

Data is an important resource for any organisation.

It is the source of information and allows an organisation to operate efficiently and effectively.

Therefore it is vital that data is managed appropriately whether it is being collected, organised, stored, communicated or deleted.

File naming strategies

In most information systems some method of naming items is required (consider the database SAC).

The main goal is to use a consistent, sensible system, especially if the project is a collaborative one.

File naming strategies

Formats:

Different operating systems have different file naming rules. You must be aware of these rules when setting your strategies. Examples include:

- ▶ Windows & Mac OS do not allow certain characters (/ , ; , ?).
- ▶ Both limit the number of characters in a file name.

File naming strategies

Conventions:

These are guidelines that should be followed when creating file (or other object) names.

A general rule is that names should be short, meaningful & consistent.

File naming strategies

Conventions:

Some examples include –

- ▶ Use numerals, not words, for numbers (sort appropriately). Use leading zeroes for better sorting).
- ▶ Avoid punctuation marks & special characters.
- ▶ Spaces can be an issue (use CamelCase).
- ▶ Use version numbers or dates.

Organising & storing data

Some points to consider:

- ▶ Organise files using hierarchical structures to improve efficiency.
- ▶ Using the cloud to store data has many advantages (p 148).
- ▶ Using metadata (data about data) assists in searching.

Organising & storing data

Some points to consider:

- ▶ Consider archiving files, although this does have some issues.
- ▶ Keep your email folders well organised.
- ▶ Using file synchronisation software (identical copies on multiple devices) ensures all versions are up to date (timely).

Identifying patterns & relationships between data

It is often difficult to recognise patterns or relationships between data.

Statistics is a tool that can be used to convert large quantities of data into small, informative, meaningful summaries.

Statistical concepts are essential for manipulating data and reaching sensible conclusions.

Some useful statistical concepts

Average

A single value that gives the most representative summary value of a range of numbers.

Different methods of averaging data include:

- ▶ Mean – the sum of the data divided by the total number of values.
- ▶ Median – the number in the middle of a range.
- ▶ Mode – the value that occurs most frequently.

Some useful statistical concepts

Significance

A significance test can help indicate when a difference may be important.

An example of a measure of significance is standard deviation.

For example a low SD shows that there is little variation in a mean value, so the value is representative and significant.

A high SD indicates that the variation between values is high so the mean may lack significance.

Some useful statistical concepts

Correlation & causality

This looks at relationships between data and what may have caused the patterns.

Correlation measures whether trends are related in any way.

Causation looks for what is responsible for the patterns.

Be careful not to just assume that a correlation is caused by one particular aspect (see examples on p 157).

Identifying patterns & relationships between data

Data visualisations

Using visual techniques (such as lines, shapes & colours) can make it easier to interpret data.

This can include graphs, charts, maps, histograms & diagrams.

The aim is to see the main aspect and trends in the data (p 157 – 159).

Identifying patterns & relationships between data

Queries & searches

These are a way of managing large data sets. The goal is to hide most of the data and highlight the interesting parts.

Sorting and filtering data can be used to achieve this goal.

Identifying patterns &
relationships between data

Conditional formatting

Allows you to change the appearance
of data automatically based on its
current value.

Can be used with database and
spreadsheet applications.

See p 164 and Wikipedia climate
data.

Digital systems

An information system is a system that converts data to information by processing it.

To achieve this function an information system is made up of a number of components which interact to create, control and communicate ideas and digital solutions.

Digital systems

These components are:

- ▶ Data
- ▶ Processes
- ▶ People
- ▶ Digital systems (hardware & software).

Digital systems

Hardware for various purposes

- ▶ Input: keyboard, mouse, trackpad, touchscreen, scanners.
- ▶ Output: monitors, printers, speakers.
- ▶ Storage: primary (RAM) & secondary (HDD, solid state drives).
- ▶ Network: ports, switches, routers, cables, WAPs.

Digital systems

Software

The programming code that controls hardware. Types include:

- ▶ Application software (RDBS)
- ▶ System software (operating systems, device drivers)
- ▶ Utility software (malware protection)

Data security

Keeping an organisations data secure is important for a number of reasons:

- ▶ Unable to operate effectively
- ▶ Loss of secrets to competitors
- ▶ Loss of reputation & trust
- ▶ Legal requirements

All of these will impact on a business's ability to be profitable.

Data security

There are many threats to an organisation's data, some deliberate, some accidental. These include:

- ▶ Employee error
- ▶ Climatic conditions
- ▶ Power issues
- ▶ Hardware or software failure
- ▶ Malicious software
- ▶ External attacks (hacking)

Data security

Physical security

This refers to keeping threats away from your data:

- ▶ Restrict physical access (locks etc)
- ▶ Surge protectors & UPS

Data security

Software security

- ▶ Usernames & passwords
- ▶ Restrict access (security levels)
- ▶ Biometric identification
- ▶ Encryption
- ▶ Firewalls
- ▶ Antivirus (malware) software
- ▶ Backing up

Data security

Keeping an organisations data secure is important for a number of reasons:

- ▶ Unable to operate effectively
- ▶ Loss of secrets to competitors
- ▶ Loss of reputation & trust
- ▶ Legal requirements

All of these will impact on a business's ability to be profitable.

Data security

There are many threats to an organisation's data, some deliberate, some accidental. These include:

- ▶ Employee error
- ▶ Climatic conditions
- ▶ Power issues
- ▶ Hardware or software failure
- ▶ Malicious software

External attacks (hacking)

Data security

Physical security

This refers to keeping threats away from your data:

- ▶ Restrict physical access (locks etc)
- ▶ Surge protectors & UPS

Data security

Software security

- ▶ Usernames & passwords
- ▶ Restrict access (security levels)
- ▶ Biometric identification
- ▶ Encryption
- ▶ Firewalls
- ▶ Antivirus (malware) software
- ▶ Backing up

